Haplotype Phasing

CSCI2820

Class 5

transcribed by derek aguiar

Haplotype Phasing algorithms

- Last class we covered
  - Clark's Algorithm
- Today we will cover
  - Haplotype Phasing by Expectation Maximization (EM)
- Later, we will cover advanced topics in the Clark algorithm
  - Algorithmic analysis, time complexity
  - Introduction to graph theory
  - How it relates to population genomics and statistical genetics including Ewen's Sampling Lemma (Formula)

Preliminaries

- Genotypes are strings of 0, 1, and 2
  - 0 and 1 denote homozygous for the major and minor allele respectively
  - 2 is the ambiguous heterozygote containing both 0 and 1 alleles at a variant
- Examples of genotypes
  - 012012 (genotype with 2 heterozygous variants)
  - 011011 (genotype of all homozygous variants)

Preliminaries

- With *k* ambiguous sites, the number of pairs of haplotype explanations is exponential in *k*
  - Ex. 012012 has haplotype explanations
    - 011011 and 010010
    - 011010 and 010011
- In general, a haplotype explanation is denoted by a pair of haplotypes (h_1,h_2)

Recall the Clark Algorithm

- Have a set of genotypes
- Look in the input set, find genotypes with unique explanation
- Add to resolved set
- Use haplotypes in resolved set to find haplotypes consistent with unresolved genotypes
- Use consistent haplotypes to phase unresolved genotypes

Clark *Method*

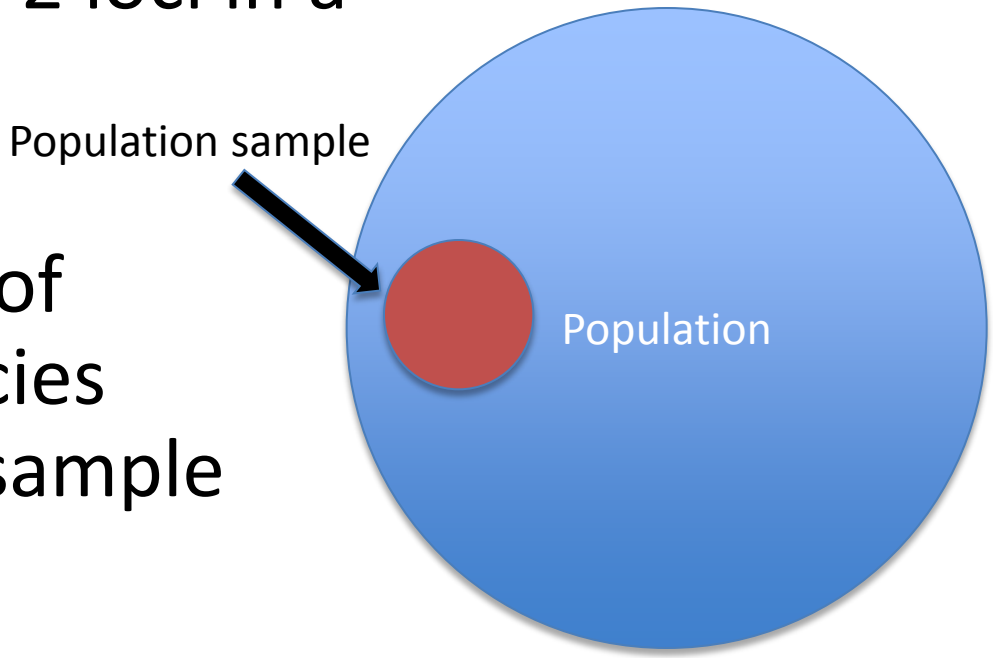
- Defining the problem
  - Input is *t* genotypes $g_1, g_2, \ldots, g_t$
  - Run clark algorithm
  - Output
- This is not a well-defined problem
- A better way to talk about the Clark algorithm is the refer to it as Clark's method because there are many ways to formulate Clark's algorithm
  - E.g. different rules to order the resolution of genotypes

Expectation Maximization Algorithm

- EM can be used for
  - Estimation of Haplotype Frequencies in a population
  - Haplotype phasing
- EM is an algorithm for obtaining a local optimum for the problem of maximum likelihood (ML)
  - the global likelihood of ML is often very difficult to compute directly

EM by example

- Problem: Consider 2 loci, with two alleles at each loci (heterozygous)
- Given: (observed) the genotype of the individuals at these 2 loci in a sample
  - g1=22 and g2=22
- Find: The estimate of haplotype frequencies at the 2 loci in the sample



EM by example

- Haplotypes are unknown, they are latent variables (hidden)
- We need to infer the latent variables (haplotypes) from the observed data (genotypes)
- In our example, there are 4 possible haplotypes 00, 01, 10, 11
- Let Θ_{00}^{t}, Θ_{01}^{t}, Θ_{10}^{t}, Θ_{11}^{t} be the haplotype frequencies in the sample at time *t*

Expanding “hidden” variables

- The hidden variables are now
  - The haplotypes
  - The frequencies of the explanations
- n individuals in the sample, the sets n_i define the set of individuals with a particular genotypic configuration
- n_A {0,1} {0,1} or in our notation 22
- n_B {0,0} {0,1} or 02
- And similarly define the counts for n_C, n_D, and n_E corresponding to genotypes 20, 00, 11 respectively
- Note: this doesn’t cover all cases of genotypes, e.g. we are missing 12 and 21, but you don’t have to include all explanations if they do not exist in the sample

EM by example

- Step from time *t* to time *t+1*
- Case (A) n_A has two explanations (00,11) or (01,10) denote this explanation Y_A
- $P(Y_A)=2\Theta_{00}^{t}\Theta_{11}^{t}+2\Theta_{01}^{t}\Theta_{10}^{t}$
- $P(00,11|Y_A)=2\Theta_{00}^{t}\Theta_{11}^{t}/(2\Theta_{00}^{t}\Theta_{11}^{t}+2\Theta_{01}^{t}\Theta_{10}^{t})$
- Case (B) n_B has one explanation (00,01)
- Case (C) n_C has one explanation (00,10)
- Likewise for case (D) and (E)

EM by example

- $n_{00}^{(t+1)}$ is the total expected number of 00 haplotypes at time t+1
- $n_{00}^{(t+1)} = n_A P((00,11)|Y_A) + n_B + n_C + 2n_D + 0n_E$
- So we update $\Theta_{00}^{t+1} = n_{00}^{t+1}/2n$

Introducing the maximum likelihood polynomial

- Genotype *g* has *r* explanations: (h_{11},h_{12}) (h_{21},h_{22}) … (h_{r1},h_{r2})
- We are going to think about *g* as follows: $(x_{11}*x_{12}+x_{21}x_{22}+\ldots+x_{r1}*x_{r2})=1$ because the x's are probabilities and should add up to 1
- The polynomial are the terms for each genotype in the sample
- The maximum likelihood polynomial maximizes over all x's

Next class

- We will describe the EM algorithm in full generality